Discrete Optimization 21 (2016) 71–84

Contents lists available at ScienceDirect

Discrete Optimization

www.elsevier.com/locate/disopt

Efficient solutions for weight-balanced partitioning problems

Steffen Borgwardt [a,*], Shmuel Onn [b]

[a] *Technische Universität München, Germany*
[b] *Technion - Israel Institute of Technology, Haifa, Israel*

ARTICLE INFO

Article history:
Received 30 March 2015
Received in revised form 23 May 2016
Accepted 1 June 2016
Available online 23 June 2016

MSC:
90C10
90C25
90C90
91C20

Keywords:
Constrained clustering
Convex maximization
Integer programming
Land consolidation

ABSTRACT

We prove polynomial-time solvability of a large class of clustering problems where a weighted set of items has to be partitioned into clusters with respect to some balancing constraints. The data points are weighted with respect to different features and the clusters adhere to given lower and upper bounds on the total weight of their points with respect to each of these features. Further the weight-contribution of a vector to a cluster can depend on the cluster it is assigned to. Our interest in these types of clustering problems is motivated by an application in land consolidation where the ability to perform this kind of balancing is crucial.

Our framework maximizes an objective function that is convex in the summed-up utility of the items in each cluster. Despite hardness of convex maximization and many related problems, for fixed dimension and number of clusters, we are able to show that our clustering model is solvable in time polynomial in the number of items if the weight-balancing restrictions are defined using vectors from a fixed, finite domain. We conclude our discussion with a new, efficient model and algorithm for land consolidation.

1. Introduction

Partitioning a set of items while respecting some constraints is a frequent task in exploratory data analysis, arising in both operations research and machine learning; see e.g. [1,2]. We consider partitioning for which the sizes of the clusters are restricted with respect to multiple criteria. There are many applications where it is necessary to adhere to given bounds on the cluster sizes.

For example, these include the modeling of polycrystals in the material sciences [3] and face recognition using meshes, where the original mesh is partitioned into parts of equal sizes to obtain an optimal running time for graph-theoretical methods that are applied to all of these parts [4]. Our interest in these types of problems comes from an application in *land consolidation*. See for example [5–7] and in particular the outreach article [8] for the impact in academia and practice.

* Corresponding author.
E-mail addresses: borgwardt@ma.tum.de (S. Borgwardt), onn@ie.technion.ac.il (S. Onn).

http://dx.doi.org/10.1016/j.disopt.2016.06.001
1572-5286/

1.1. Land consolidation

The farmers of many agricultural and private forest regions in Bavaria and Middle Germany own a large number of small lots that are scattered over the whole region. The reasons are strict heritage laws and a frequent change of ownership. There is significant overhead driving and an unnecessarily high cost of cultivation. In such a situation, a land consolidation process may be initiated by the state to improve on the cost-effective structure of the region. Voluntary land exchanges (by means of lend-lease agreements) are a popular method for such a process: The existing lots are kept without changes and the rights of cultivation are redistributed among the farmers of the region.

This corresponds to a combinatorial redistribution of the lots and can be modeled as a clustering problem where each lot is an item and each farmer is a cluster [7]. The main goal is to create large connected pieces of land for each farmer. One way to do so is to represent the lots by their midpoints in the Euclidean plane and to use the geographical locations of the farmsteads of each farmer as a set of sites. Then one performs a weight-balanced least-squares assignment of the lots to these sites [5,9,10]. The result is a redistribution where the farmers' lots lie close to their farmsteads. As a positive side effect, many of a farmer's lots are connected and can be cultivated together.

The lots differ in several features like their size, quality of soil, shape, and attached subsidies, and some of these features are even different for each farmer. For example one farmer may be eligible for subsidies if they cultivate a given lot, while another farmer may not. In such a situation, the lot is more valuable for the first farmer. Here a natural constraint is that – after the redistribution – each farmer should have lots that (approximately) sum up to the farmer's original total with respect to each feature.

Of course, partitioning a weighted set of items into clusters of prescribed sizes (*weight-balanced partitioning*) is readily seen to be NP-hard, even for just two clusters and each item having just a single weight that is uniform for both clusters: deciding whether there is such a partition is at least as hard as the *Subset Sum* problem. The methods in the literature [5,8–10] approach this intrinsic hardness by solving least-squares problems by an LP relaxation and rounding. The model in [5] performed particularly well in practice.

In this paper, we will present a general clustering framework that, when applied to land consolidation, improves on this model by dealing with its biggest shortcoming: The model in [5] (and in fact [8–10], too) is not able to balance the weights of clusters with respect to multiple features of the lots at the same time. Instead the redistribution of cultivation rights is done with respect to a single 'value' of a lot, an aggregation of all its properties (and this value is the same for each farmer).

The farmers will not accept a large deviation with respect to any of the features of their total lots. They will only participate in the redistribution if they do not lose a significant area of land, do not lose a lot in quality of soil, and do not lose much of their subsidies at the same time. But the models in the literature will return 'optimal' solutions for which not even the aggregated value necessarily is within the specified bounds (e.g. 3%) from the original. Further, even if a farmer's aggregated deviation is small, they may have received more land than before, but of much lower quality—which they will not accept. These are intrinsic weaknesses coming from the relaxation and rounding that are performed in [5,8–10]. In the practical implementation, this meant that a lot of the work still had to be done 'by hand'.

However, the previous methods do not use all of the favorable properties of the input data for agricultural regions. The data typically falls into only a fixed number of categories, which just comes from the way the features are measured in practice. For example, especially large slots (>5 ha) are not traded at all. Further, one does not distinguish between lot sizes that differ by less than a tenth of a hectare, so that one obtains a finite domain of lot sizes. The quality of soil is measured with a number between 1 and 100, which is a finite domain itself. But in fact in a single agricultural region it is rare to have more than five different values within this range. The same happens for the subsidies attached to lots and other measures. With this